eip ESCUELA INTERNACIONAL POSGRADOS

Máster en Programación avanzada en Python para Hacking, BigData y Machine Learning

Fundamentos de IA y Machine Learning

LECCIÓN 01

Introducción al Aprendizaje Automático

ÍNDICE

INTRODUCCIÓN

El Aprendizaje Automático engloba una gran cantidad de conceptos y procedimientos que son difíciles de entender sin una perspectiva introductoria a esta rama de la ciencia. En esta lección se pretende abordar estos conceptos introductorios para el posterior análisis de los datos y una primera toma de contacto con la visualización de estos.

OBJETIVOS

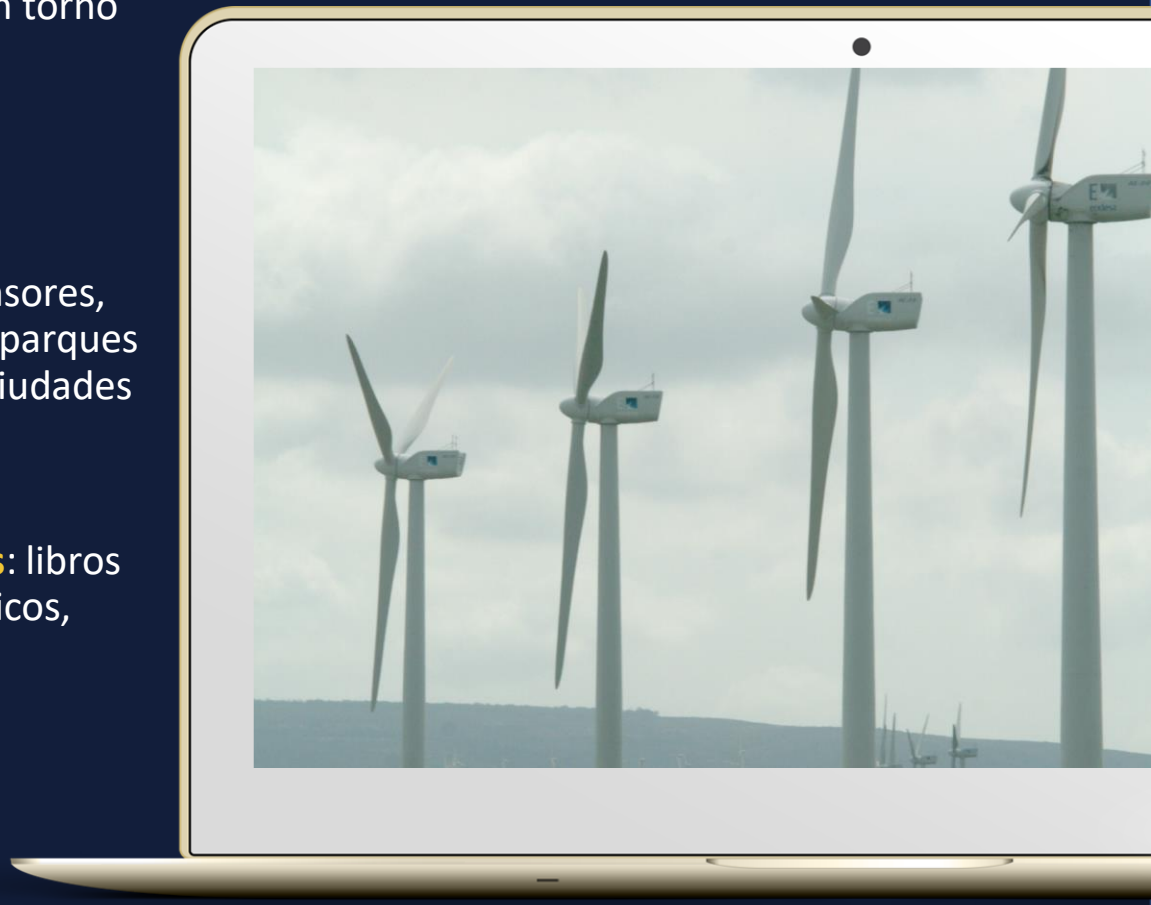
Al finalizar esta lección serás capaz de:

1. Saber que es la ciencia de datos y el aprendizaje automático.
2. Entender la diferencia entre los distintos problemas existentes.
3. Comprender las distintas etapas en las que se divide un problema de *Machine Learning*.
4. Introducir los primeros conceptos del preprocesamiento de los datos.
5. Visualizar los datos de forma gráfica

1.1. Motivación

Nuestro mundo gira cada vez más en torno a los datos:

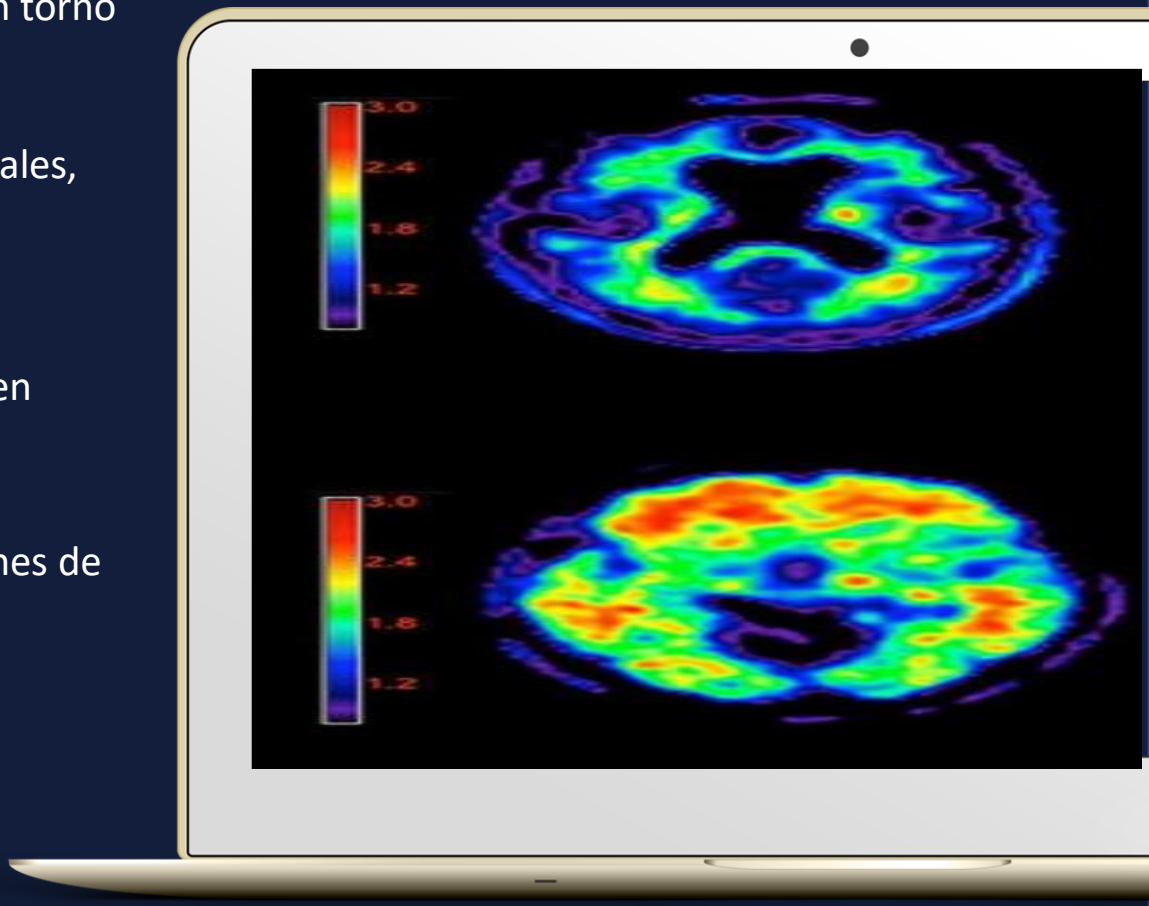
- **Ciencia**: astronomía, genómica, medioambiente,…
- **Industria y Energía**: redes de sensores, Internet de las cosas, gestión de parques eólicos, previsión de demanda, ciudades inteligentes,…
- **Ciencias sociales y humanidades**: libros digitalizados, documentos históricos, datos sociales,…



1.1. Motivación

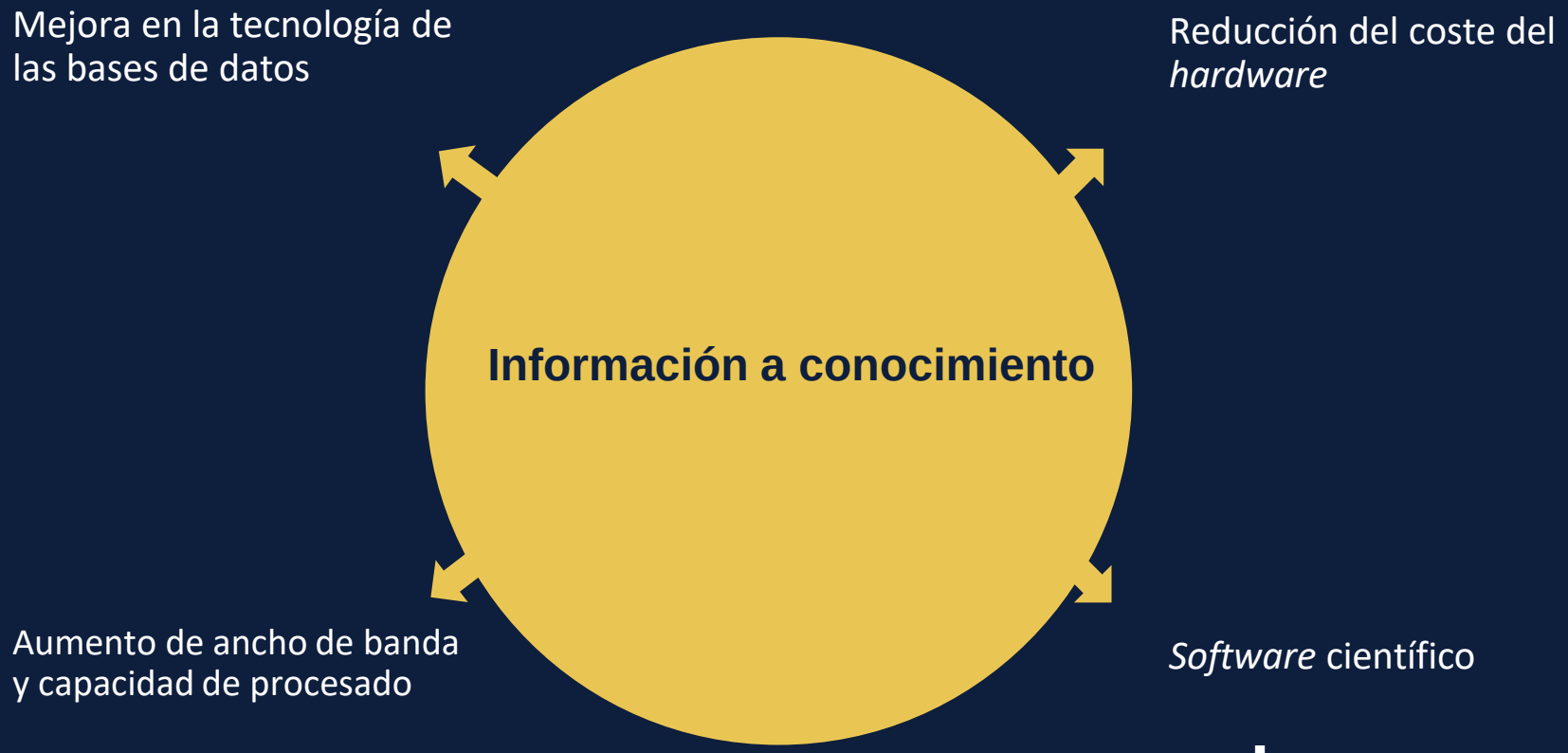
Nuestro mundo gira cada vez más en torno a los datos:

- **Entretenimiento**: sistemas de recomendación, contenidos digitales, búsquedas multimedia…
- **Medicina**: examen de imágenes médicas, previsión de demanda en hospitales, sistemas expertos…
- **Financias y negocios**: transacciones de mercados automatizadas



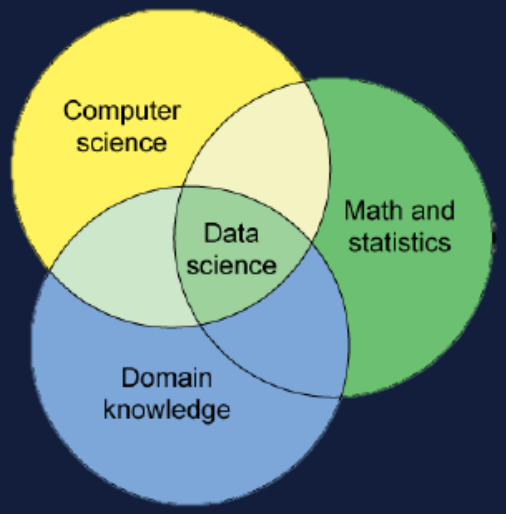
1.1. Motivación

Se ha almacenado una gran cantidad de datos que no han podido ser procesados hasta ahora.



1.2. Ciencia de datos

Ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos.

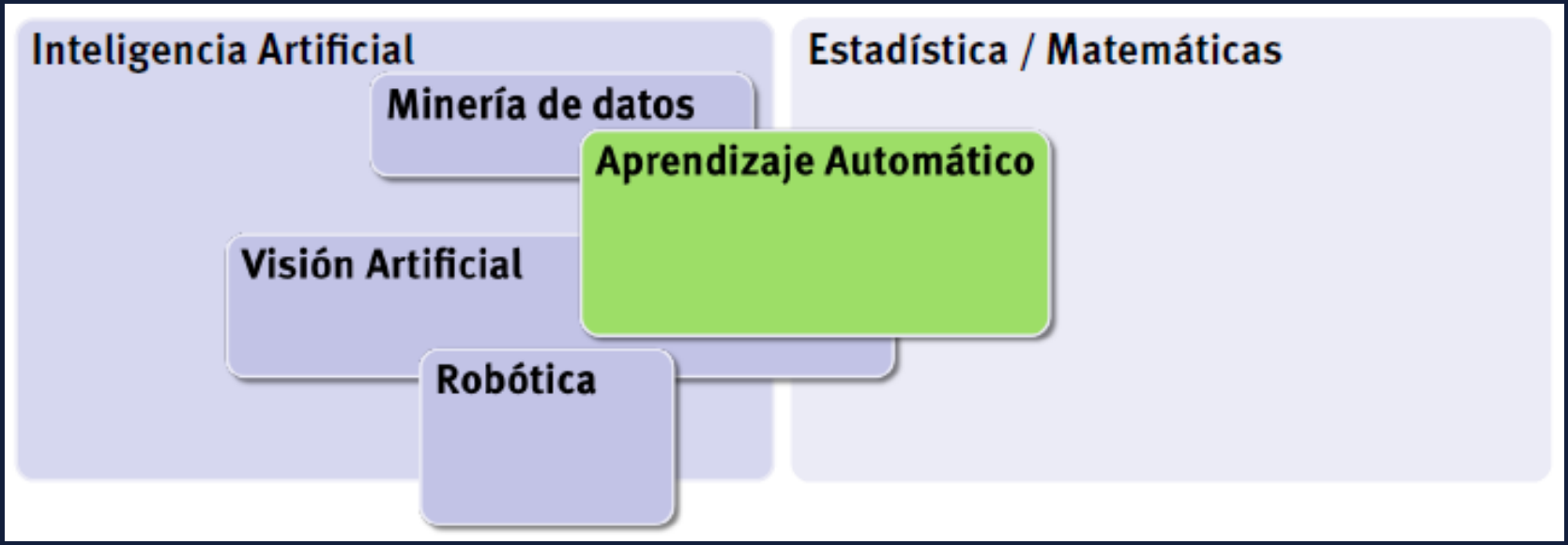


Científico de datos

Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos. De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco.”

2. Aprendizaje automático

Campo de estudio que proporciona a los ordenadores la capacidad de aprender de los datos sin haber sido explícitamente programados para ello.



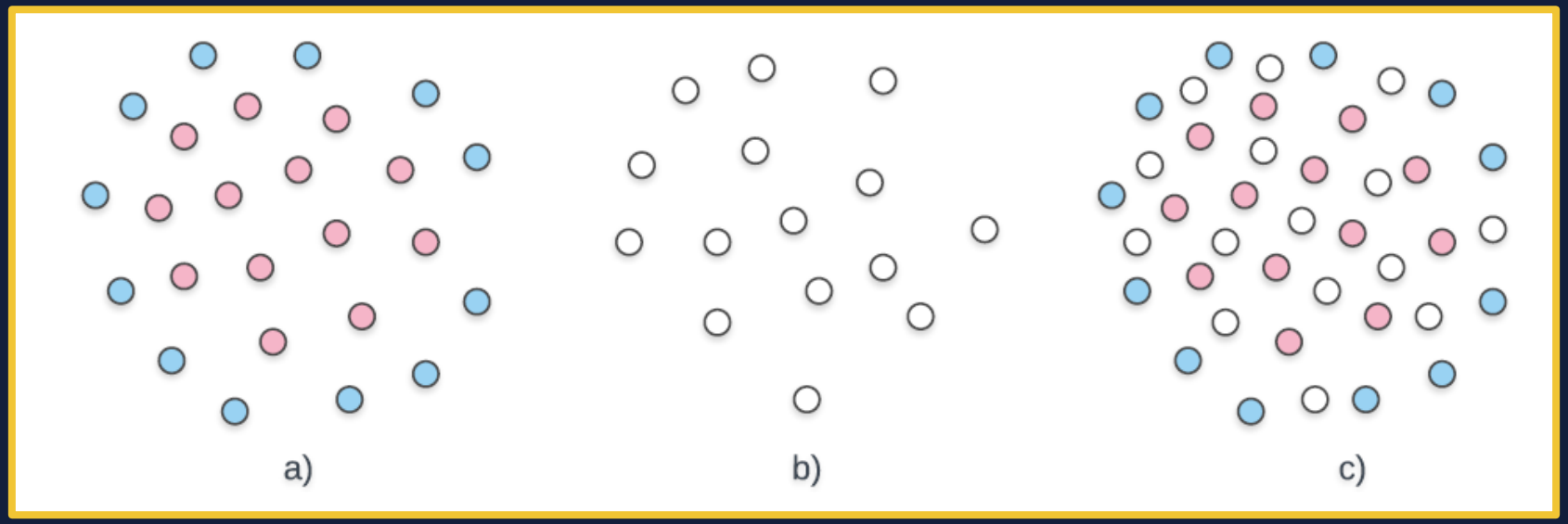
2.1. Presentación de los datos

- Fila: patrón, instancia u observación.
- Columna: característica, atributo o variable.
- Puede existir una variable objetivo, de salida o dependiente, que será el valor que se intente predecir.

Edad	Fuma	Deporte	Comida saludable	Enfermedad
28	No	Sí	Sí	**No**
16	No	No	No	**No**
45	Sí	Sí	No	**Sí**
65	Sí	No	No	**Sí**

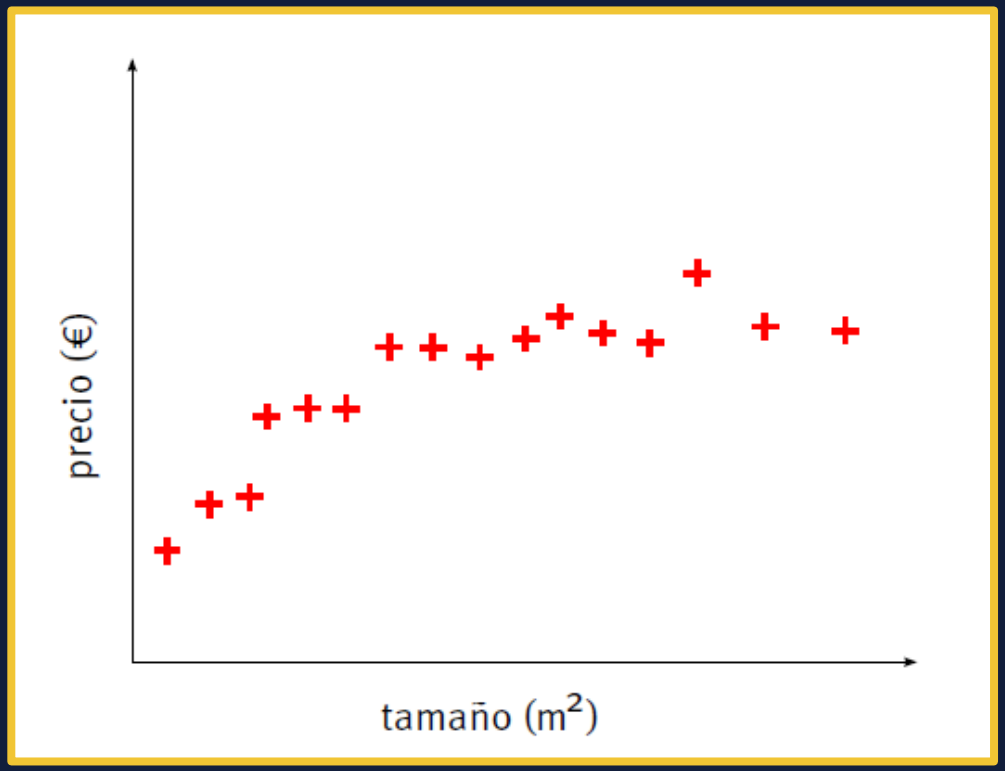
2.2. Clasificación de los métodos en aprendizaje automático

a) Aprendizaje supervisado
b) Aprendizaje no supervisado
c) Aprendizaje semisupervisado



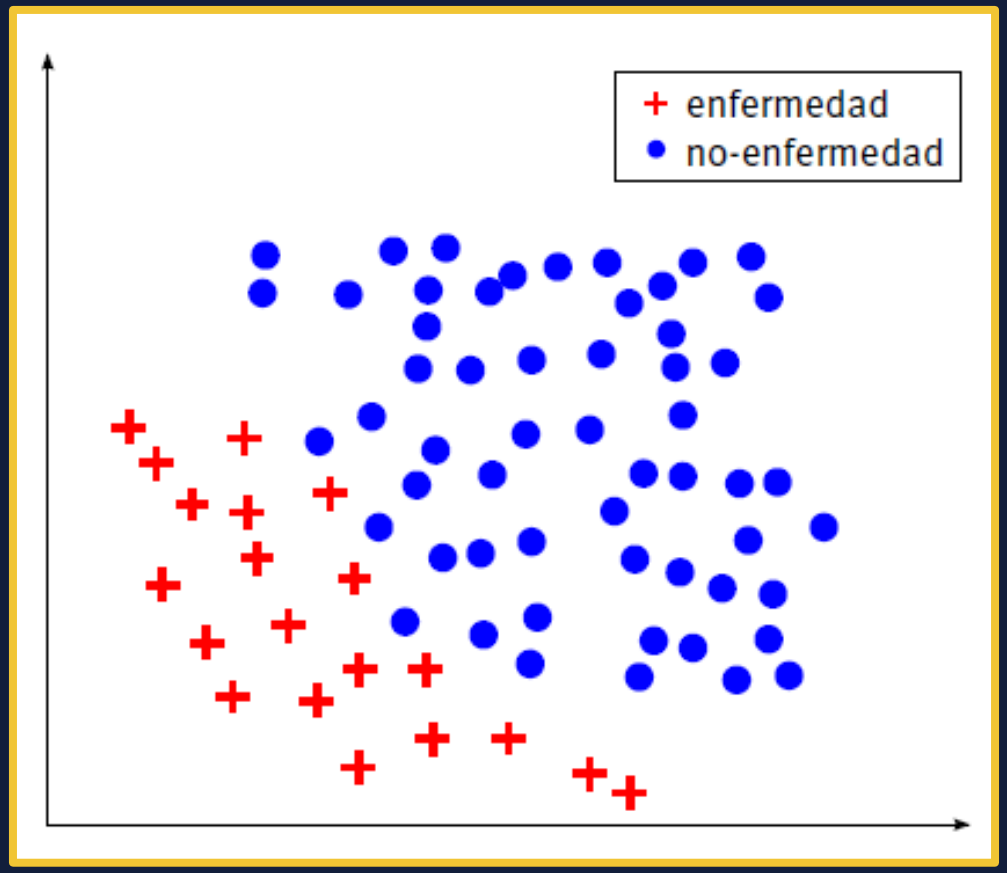
2.2.1. Aprendizaje supervisado

Regresión: la variable de salida Y es un valor real.



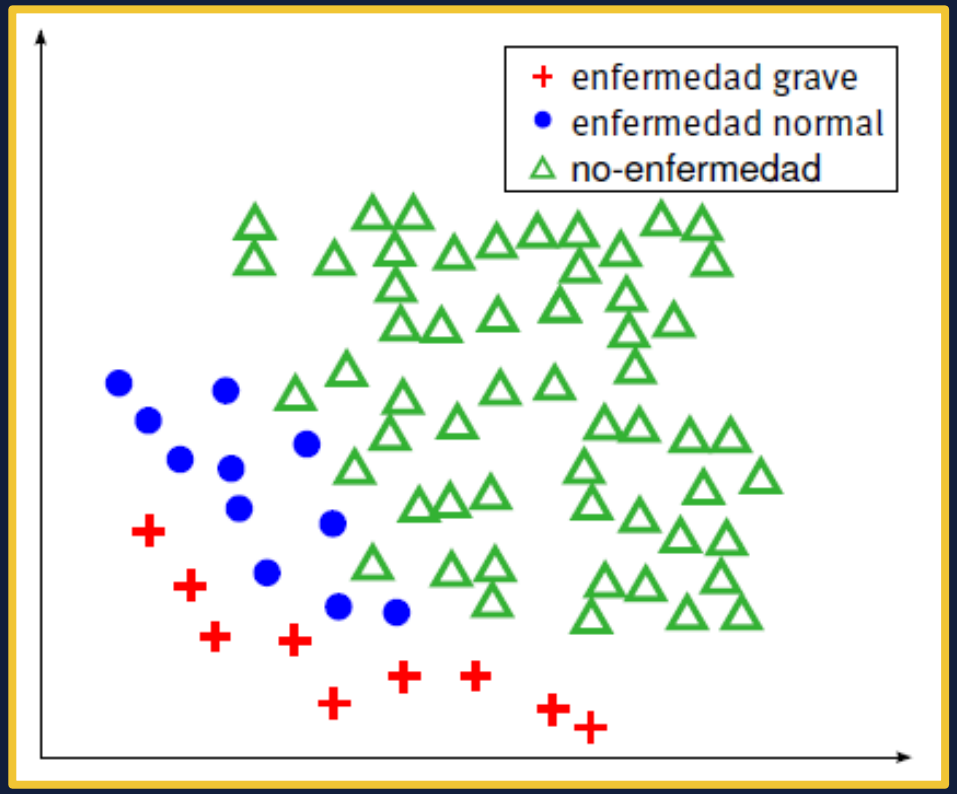
2.2.1. Aprendizaje supervisado

Clasificación: la variable de salida Y es un valor categórico, discreto o nominal.

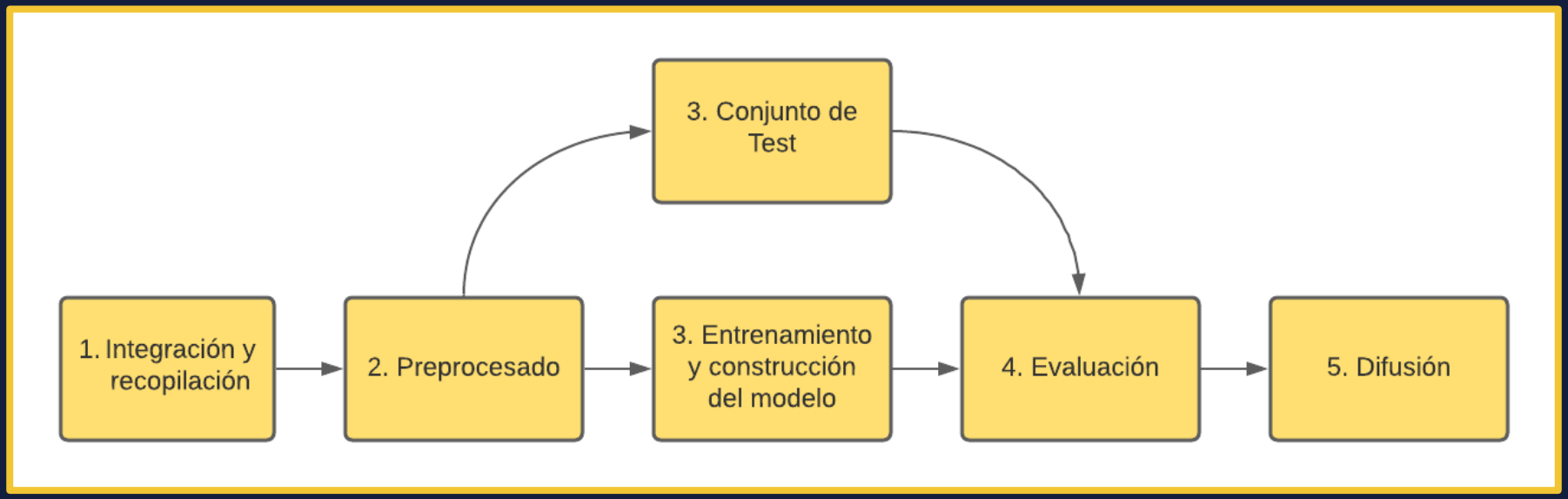


2.2.1. Aprendizaje supervisado

Clasificación o regresión ordinal: la variable de salida Y es un valor categórico, discreto o nominal, pero además existe un orden entre las clases no cuantificable.



2.3. Fases del proceso de construcción de un modelo de ML



3. Preprocesado de los datos

- A menudo, los resultados van a depender más de la calidad de los datos en relación al problema que de la parte de generación del modelo en sí.

- Es frecuente hablar de ruido en los datos, de relevancia de variables… siempre con respecto a un objetivo. Una variable puede ser ruido en un problema o información muy útil en otro.

- Disponemos de potentes modelos de aprendizaje automático no lineales capaces de ajustarse a datos complejos y de alta dimensionalidad. Por ejemplo, las redes neuronales son considerados aproximadores universales.

3.1. Selección de variables o características

Para obtener una buena base de datos es necesario aplicar:

- **Extracción de características:** determinar que variables necesitamos para la caracterización de un problema. Por ejemplo, al procesar datos multimedia se extraen características que permiten construir vectores de tamaño fijo necesarios para los modelos.

- **Selección de características:** este proceso consiste en descartar aquellas características que no son necesarias en nuestro problema. Por ejemplo, variables con un valor constante, variables que carecen de significado en un problema (color de ojos para determinar si una persona tiene riesgo de sufrir un ataque cardiaco), etc.

3.2. Limpieza de datos

Se realizan operaciones sobre los datos, ya sean variables (columnas) o patrones (filas).

Recuperación de valores perdidos

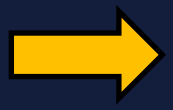
Tratamiento de valores anómalos (*outliers*) o inconsistencias

Tratamiento del ruido

3.3. Transformación de datos

Binarización de variables categóricas

Animal
Perro
Gato
Conejo
Perro



Perro	Gato	Conejo
1	0	0
0	1	0
0	0	1
1	0	0

3.3. Transformación de datos

Escalado de variables

Normalización N(0,1)

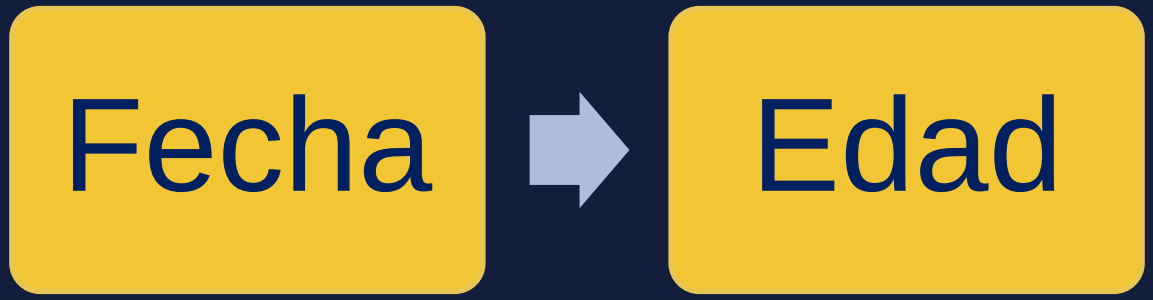
$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$$

Escalado [0, 1]

$$x'_{i,j} = \frac{x_{i,j} - min(x_j)}{max(x_j) - min(x_j)}$$

3.3. Transformación de datos

Transformar fechas



3.4. Reducción de la dimensionalidad

Es distinto a la selección de características. Se basa en emplear técnicas como el análisis de componentes principales (PCA) para obtener combinaciones lineales de variables que reduzcan la dimensión de los datos.

3.5. Tratamiento del desbalanceo

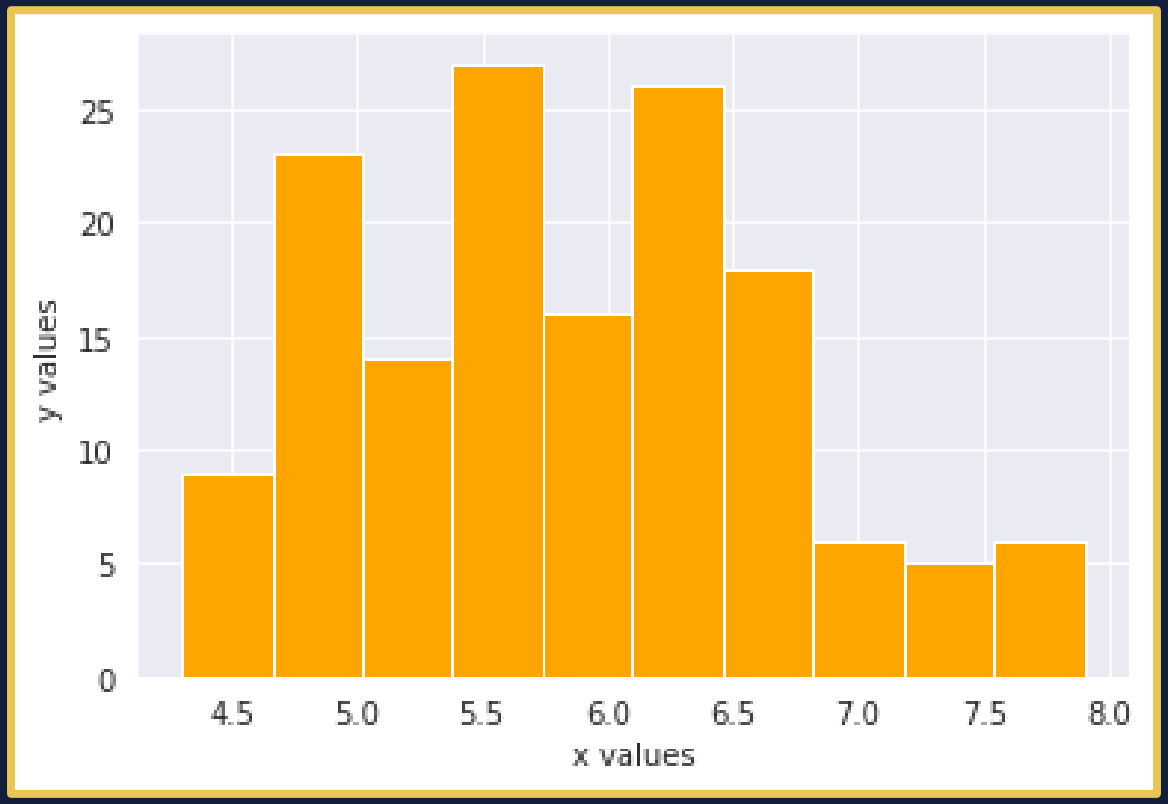
Puede ocurrir que, en problemas supervisados de clasificación, exista un número de patrones distinto por cada clase. Esto da lugar al desbalanceo, es decir, que el número de patrones de una clase sea muy grande con respecto a otra. En este sentido, los algoritmos de aprendizaje automático tienden a "aprender" a clasificar bien las clases mayoritarias. Para evitar este problema tenemos dos alternativas:

- **Técnicas de *oversampling*:** se generan patrones sintéticos en aquellas clases con menor número de patrones.

- **Técnicas de *undersampling*:** se reduce el número de patrones de las clases mayoritarias.

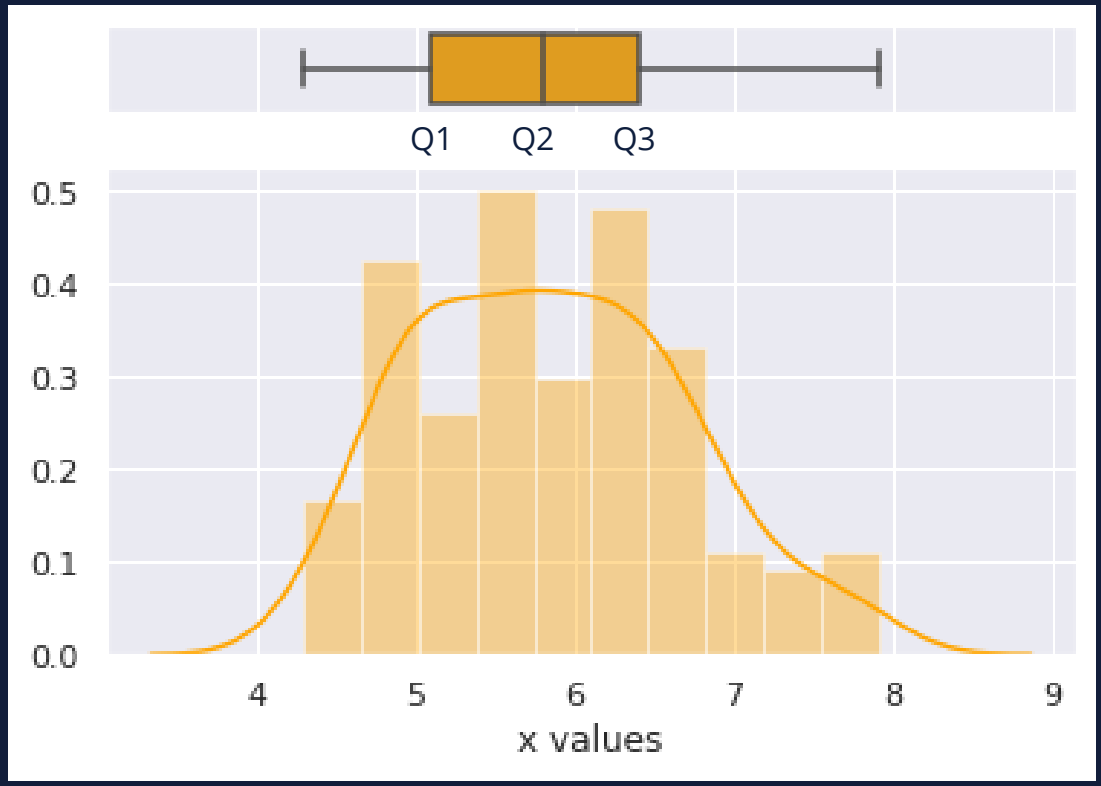
4.1. Distribución de los datos

Histograma



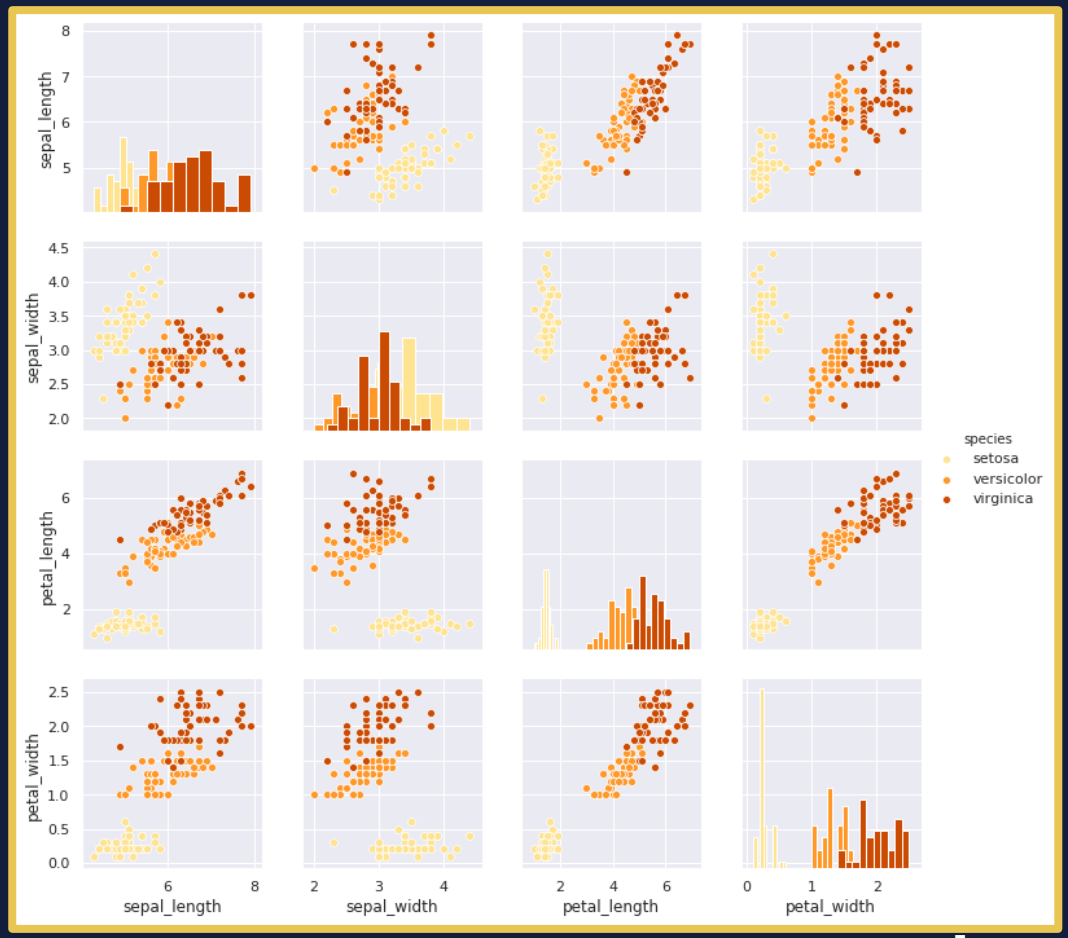
4.1. Distribución de los datos

Boxplot o diagrama de cajas



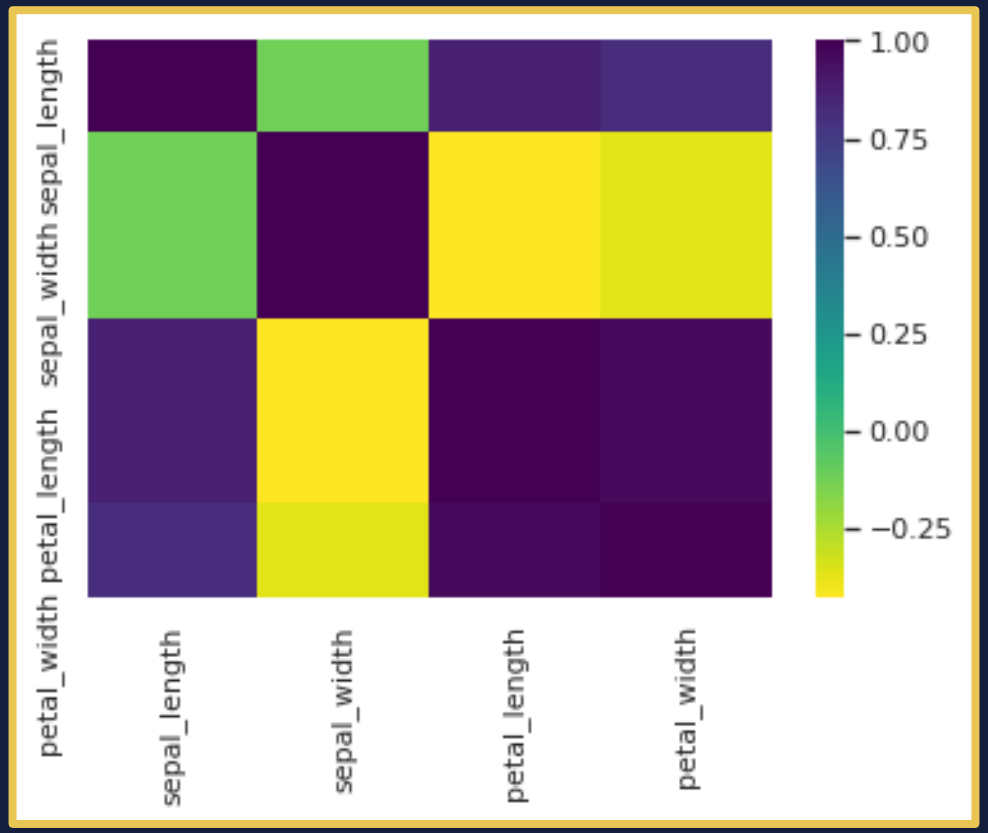
4.2. Correlación entre variables

Scatter-plot o diagrama de dispersión



4.2. Correlación entre variables

Heatmap o mapa de calor



MUCHAS GRACIAS POR SU ATENCIÓN

eip ESCUELA INTERNACIONAL POSGRADOS


antoniomdr18@gmail.com


Antonio Manuel Durán Rosal
https://www.linkedin.com/in/antonio-manuel-dur%C3%A1n-rosal-99ba92a5/


twitter.com/eiposgrados


facebook.com/eiposgrados


instagram.com/eiposgrados